Topic: Introduction in Spoken Language

Part I
Spoken Language Systems
Spoken Language Structure
Speech Perception
Speech Production

Part II
Phonetics and Phonology
Syllables and Words
Syntax and Semantics
Transcription of Spoken Speech

Reference: Huang et al. Chapter 1 and 2; Waibel/Lee Chapter 9

April 24, 2012
Overview (I)

- Spoken Language Systems
  - Motivation
  - Structure
  - Research areas and Foundations
  - Some Historic Milestones
  - Progress in Speech Recognition
  - (D)ARPA Themes
  - Humans vs Machines on similar tasks
  - Taxonomy of Speech Recognition
  - Why Is Speech Recognition Difficult?
  - What factors influence the difficulty?
  - So why Is it Easy for Humans?
Overview (II)

• **Spoken Language Structure**
  - The Speech Chain
  - The Structure of Spoken Language
  - Sound
    - What Is Sound?
    - Wavelength, Amplitude, Frequency of a Wave
    - Measuring the Intensity of Sound
    - Examples for Sound Levels in Decibel
    - Pick up the Sound Pressure Wave (Microphones)

• **Speech Perception**
  - Physiology of the Ear
  - Biology of Perception
  - Physical versus Perceptual Attributes
  - Non-uniform Equal Loudness
Overview (III)

- **Speech Production**
  - Human Speech Production
  - Fundamental Frequency
  - Voiced vs. Unvoiced Sounds
  - Articulatory Properties
  - Speech Sound Names
  - Consonants versus Vowels
Overview

- Spoken Language Systems
  - Motivation
  - Structure
  - Research areas and Foundations
  - Some Historic Milestones
  - Progress in Speech Recognition
  - (D)ARPA Themes
  - Humans vs Machines on similar tasks
  - Taxonomy of Speech Recognition
  - Why Is Speech Recognition Difficult?
  - What factors influence the difficulty?
  - So why Is it Easy for Humans?
Motivation – Why Speech?

• “One cannot not communicate.” (Paul Watzlawick)

• Human-Human Communication: Why Speech?
• Speech is the most natural form of human communication
  • From prehistory to the new media of the future, speech communication has been the dominant mode of human social bonding and information exchange
  • The spoken word is now extended through technological mediation such as telephony, movies, radio, TV, Internet
  • … so far nothing can truly replace a spoken conversation (think about misunderstandings resulting from pure text messages, such as email, SMS)
• Digital Campfire – information exchange in media communities – also preferred in speech and pictures
Does this hold for Human-Computer Interaction?

<table>
<thead>
<tr>
<th>Communication</th>
<th>Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong></td>
<td></td>
</tr>
<tr>
<td>- Keyboard, mouse, touch screen</td>
<td>tactile</td>
</tr>
<tr>
<td>- Microphone</td>
<td>audio</td>
</tr>
<tr>
<td>- Scanner</td>
<td>visual</td>
</tr>
<tr>
<td>- Camera, Eye tracking, Gaze tracking</td>
<td>visual</td>
</tr>
<tr>
<td><strong>Output</strong></td>
<td></td>
</tr>
<tr>
<td>- Display</td>
<td>visual</td>
</tr>
<tr>
<td>- Loudspeaker</td>
<td>audio</td>
</tr>
</tbody>
</table>

Today’s computer still lack the human ability to speak, listen, understand and learn. **WHY?**
## Comparison of Input Modes

<table>
<thead>
<tr>
<th>Mode</th>
<th>Speed</th>
<th>Reliability</th>
<th>Devices</th>
<th>Practice</th>
<th>Other Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handwriting</td>
<td>200-500</td>
<td>Recognition errors</td>
<td>Tabloid, scanner, BIG</td>
<td>Literacy</td>
<td>Hands and eyes busy</td>
</tr>
<tr>
<td>Typing</td>
<td>200-1000</td>
<td>~100% Typos</td>
<td>keyboard, BIG</td>
<td>Literacy, yes for fast input</td>
<td>Hands and eyes busy</td>
</tr>
<tr>
<td>Speech</td>
<td>up to 4000</td>
<td>Recognition errors</td>
<td>Micro, SMALL</td>
<td>NO</td>
<td>Hands and eyes free</td>
</tr>
</tbody>
</table>
A Spoken Language System – Structure

![Diagram showing the structure of a spoken language system with the following components: Database, Application, Discourse Analysis, Dialog Manager, Dialog Strategy, Speech Recognition, Sentence Interpretation, Response Generation, Synthesis, Speech, Text, Meaning, Text, Speech.]
Spoken Language Systems

• A spoken language system needs to have both: **Speech Recognition** and **Speech Synthesis**

• But this is NOT sufficient in order to build a useful spoken language system

• It also requires an **Understanding and Dialog Component** to manage interactions with the user

• **Domain knowledge** must be provided to guide the system’s interpretation of speech and to determine the appropriate action

• All these components have significant challenges such as:
  – Robustness (e.g. noise)
  – Flexibility (e.g. portability to new domains, languages)
  – Ease of Integration
  – Engineering Efficiency (e.g. real time)
Research areas and Foundations

• Spoken language processing is a diverse field that relies on knowledge of language at the levels of:
  – signal processing
  – acoustics
  – phonology
  – phonetics
  – syntax
  – semantics
  – pragmatics and discourse

• Foundations of spoken language processing lie in the field of:
  – computer science
  – electrical engineering
  – linguistics
  – psychology
### Some Historic Milestones

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1939</td>
<td>VOiCe Operated CODER (Vocoder) by Dudley (Bell Labs)</td>
</tr>
<tr>
<td>1946</td>
<td>&quot;Visible Speech&quot; by Bell (to teach deaf)</td>
</tr>
<tr>
<td>1965</td>
<td>Paper by Cooley &amp; Tukey: &quot;The Fast Fourier Transform&quot;</td>
</tr>
<tr>
<td>1968</td>
<td>Dynamic Time Warp (DTW) for speech recognition by Vintsyuk</td>
</tr>
<tr>
<td>1971</td>
<td>DARPA starts ambitious speech understanding project (SUR)</td>
</tr>
<tr>
<td>1975</td>
<td>Statistical Models (HMMs) first proposed by J. Baker at CMU</td>
</tr>
<tr>
<td></td>
<td>(ASR-History at CMU see <a href="http://www-2.cs.cmu.edu/~msiegler/ASR/futureofcmu-final.html">http://www-2.cs.cmu.edu/~msiegler/ASR/futureofcmu-final.html</a>)</td>
</tr>
<tr>
<td>1985</td>
<td>Introduction of context-dependent models (Kai-Fu Lee)</td>
</tr>
<tr>
<td>1988</td>
<td>Speaker-independent continuous speech (&gt;1000 words)</td>
</tr>
<tr>
<td>1992</td>
<td>Large vocabulary (isolated) word dictation</td>
</tr>
<tr>
<td>1995</td>
<td>Speaker-independent continuous speech (&gt;60,000 words)</td>
</tr>
<tr>
<td>1997</td>
<td>Commercially available LVCSR &gt;60,000 words (IBM, Dragon)</td>
</tr>
<tr>
<td>2000</td>
<td>Speech-to-speech translation for compact domains (Verbmobil)</td>
</tr>
<tr>
<td>2002</td>
<td>General English recognition in noisy environment (DARPA)</td>
</tr>
<tr>
<td>2004</td>
<td>Speech Translation on a PDA (Transtac – Darpa)</td>
</tr>
<tr>
<td>2005</td>
<td>GALE (Global Autonomous Language Exploitation), domain unlimited speech recognition, translation, and IR in multiple languages (English, Arabic, Chinese, news and conversation, &gt;600k vocabulary)</td>
</tr>
</tbody>
</table>
### Some Historic Milestones – The Beginning

The DARPA SUR Project (1971 - 1976)

DARPA = Defense Advanced Research Projects Agency

SUR = Speech Understanding Research

<table>
<thead>
<tr>
<th>anticipated goal by ARPA (1971)</th>
<th>HARPY (1976)</th>
</tr>
</thead>
<tbody>
<tr>
<td>connected speech</td>
<td>yes</td>
</tr>
<tr>
<td>many speakers</td>
<td>5</td>
</tr>
<tr>
<td>little adaptation</td>
<td>20 sentences/speaker</td>
</tr>
<tr>
<td>1000 words</td>
<td>1011</td>
</tr>
<tr>
<td>artificial grammar</td>
<td>perplexity 33</td>
</tr>
<tr>
<td>Word error rate</td>
<td>5%</td>
</tr>
<tr>
<td>fast (few times real time)</td>
<td>RTF 80</td>
</tr>
<tr>
<td>on 100 MIPS machine</td>
<td>on 4 MIPS (PDP-10)</td>
</tr>
</tbody>
</table>
Progress in Speech Recognition

Continual Progress in Speech Recognition at (D)ARPA Evaluations

Word Error Rates

- READ SPEECH: 1000 words
- READ SPEECH: 5000 words
- READ SPEECH: 20000 words
- READ SPEECH: 60000 words
- CONVERSATIONAL SPEECH (telephone line)
- Broadcast News
- Unlimited vocabulary

Languages:
- English (EN)
- Chinese (CH)
- Arabic (AR)
- Bengali (BN)

670,000 words
Since 2002: Multilinguality (triggered by 9/11)
**Humans vs Machines on similar tasks (2005)**

<table>
<thead>
<tr>
<th>Tasks (English)</th>
<th>Vocabulary</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Humans</td>
</tr>
<tr>
<td>Connected Digits</td>
<td>10</td>
<td>0.009%</td>
</tr>
<tr>
<td>Alphabet Letters</td>
<td>26</td>
<td>1%</td>
</tr>
<tr>
<td>Read speech (WSJ)</td>
<td>5,000</td>
<td>0.9%</td>
</tr>
<tr>
<td>WSJ noise (10db)</td>
<td>5,000</td>
<td>1.1%</td>
</tr>
<tr>
<td>Conversational Telephone Task</td>
<td>25,000</td>
<td>3.8%</td>
</tr>
<tr>
<td>Broadcast News (04)</td>
<td>100,000</td>
<td>3% (transcriber)</td>
</tr>
<tr>
<td>Broadcast Convst (noise + crosstalk)</td>
<td>100,000</td>
<td>4%</td>
</tr>
<tr>
<td>Clean speech based on 3-gram</td>
<td>20,000</td>
<td>7.6%</td>
</tr>
</tbody>
</table>

1) Humans at least 5 times better than machines, far more robust in noise and conv!
2) Same syntactic and semantic model > the difference disappears (Microsoft, 2001)
# Taxonomy of Speech Recognition

<table>
<thead>
<tr>
<th>TYPE OF SPEECH</th>
<th>SPEAKER-DEPENDENCY</th>
<th>DISCOURSE</th>
<th>SPEAKER BEHAVIOR</th>
<th>SIGNAL QUALITY</th>
<th>UTTERANCE TYPES</th>
<th>COMMUNICATION</th>
<th>VOCABULARY</th>
</tr>
</thead>
<tbody>
<tr>
<td>single commands, highly formalized or artificial languages, written/spoken, read/dictated/spontaneous/conversational, planned/unplanned, perplexity</td>
<td>single speaker (&lt;em&gt;speaker-depend.&lt;/em&gt;) fixed (small) group fixed type (gender/dialect) any speaker (&lt;em&gt;speaker-independ.&lt;/em&gt;) adaptation (&lt;em&gt;speaker-adaptive&lt;/em&gt;)</td>
<td>small: home banking, operate coffee maker simple: flight booking (ATIS), train schedule medium: travel planning (VERBMOBIL) large: patent search, business dictation unlimited: lectures, broadcast news Utopian?: human-human, general speech</td>
<td>cooperative, disciplined familiarity with system user status, disposition (stressed, tired, sad)</td>
<td>bandwidth (e.g. low on phone line) acoustic environment (office, street) noise (SNR, crosstalk, ...)</td>
<td>isolated words/phrases connected/&lt;em&gt;continuous&lt;/em&gt;</td>
<td>menus, one-way, dialog, monolog, meeting, talk, human translation</td>
<td>size, confusability, difficulty, named entities, unknown words</td>
</tr>
</tbody>
</table>
Why Is Speech Recognition Difficult?

- Speech recognition is some kind of pattern recognition:
  - Computer measures air-pressure,
  - Then interprets measurements

- Speech from the computers point of view:

- So why is speech recognition so difficult?
  - What factors influence the difficulty?
  - And why is speech recognition so easy for humans?
<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>written text:</td>
<td>Why is speech Recognition so Difficult?</td>
</tr>
<tr>
<td>spontaneous:</td>
<td>why's speech recognition so difficult</td>
</tr>
<tr>
<td>continuous:</td>
<td>whysspeechrecognitionsodifficult</td>
</tr>
<tr>
<td>pronunciation:</td>
<td>whazbeechregnizhnsadifcld</td>
</tr>
<tr>
<td>acoustic variability:</td>
<td><img src="image" alt="whazbeechregnizhnsadifcld" /></td>
</tr>
<tr>
<td>noise:</td>
<td><img src="image" alt="noise" /></td>
</tr>
<tr>
<td>cocktail party-effect:</td>
<td><img src="image" alt="cocktail party-effect" /></td>
</tr>
</tbody>
</table>
## What factors influence the difficulty?

### Complexity
- **Amount of data:** Typically 32000 bytes per second (16kHz)
- **Class inventory:** 50 phonemes, 5000 sounds, 100,000 words
- **Combinatorial explosion:** Exponential growth of possible sentences

### Segmentation
- **Our perception:** Phones, syllables, words, sentences
- **Actually there are:** No boundary markers, continuous flow of samples

### Variability
- **Speaker:** Anatomy of vocal tract, speed, loudness, acoustic stress, mood, dialect, speaking style, context
- **Channel, environment:** Noise, microphones, channel conditions

### Ambiguity
- **Homophones:** Two vs. too,
- **Word Boundaries:** Interface vs. in her face,
- **Semantics:** He saw the Grand Canyon flying to New York,
- **Pragmatics:** Time flies like an arrow.
"The main prerequisite of the uniquely human communication is that speaker and listener must have a common understanding that out of all possible sounds man can produce and hear, only a few have linguistic significance."

Interesting article by Olli Aaltonen & Esa Uusipaikka:
Why Speaking Is so Easy? – Because Talking Is Like Walking with a Mouth
(online at http://csl.anthropomatik.kit.edu/downloads/vorlesungsinhalte/MMMK-Olli_Esa.WhyIsSpeakingSoEasy.pdf)

• Important feature of speech perception:
  We hear sounds either as speech or non-speech
• Once defined as speech we hear them a sequence of vowels and consonants not as buzzes and hisses
• We segment into words on the fly, abstract away from sound variability
• We use an enormous database of background knowledge:
  – phonotactics
  – morphology
  – syntax
  – semantics
  – pragmatic knowledge
Overview

• Spoken Language Structure
  – The Speech Chain
  – The Structure of Spoken Language
  – Sound
    – What Is Sound?
    – Wavelength, Amplitude, Frequency of a Wave
    – Measuring the Intensity of Sound
    – Examples for Sound Levels in Decibel
    – Pick up the Sound Pressure Wave (Microphones)

• Speech Perception
  – Physiology of the Ear
  – Biology of Perception
  – Physical versus Perceptual Attributes
  – Non-uniform Equal Loudness
The Speech Chain (Human to Human)

A: Neuro-physiological process in the speaker's brain
B: Electrical process in the *afferent* nerves (impulses from the central nervous system)
C: Resulting position and movement of articulatory apparatus
D: Acoustical production of acoustic speech signal in vocal tract
E: Acoustical transmission of the speech signal
The Speech Chain (Human to Human)

**F**: Mechanical process in the middle ear, hydro-mechanical process in the inner ear

**G**: Electrical signals on the *afferent* nerves (impulses *to* the central nervous system)

**H**: Neuro-physiological process in the listener’s brain

**I**: Acoustic feedback to the speaker’s ear
The Speech Chain (Human and Computer)

Speech Generation

- Message Formulation
- Language System
- Neuromuscular Mapping
- Vocal Tract System

Speech Understanding

- Message Comprehension
- Language System
- Neural Transduction
- Cochlea Motion

Speech Generation

- Application semantics, actions
- Phonemes, words, Prosody
- Feature extraction
- Articulatory parameter

Speech analysis
Introduction in Spoken Language

The Speech Chain (Human and Computer)

Speech Generation

Vocal Tract System

Neuromuscular Mapping

Language System

Message Formulation

A semantic message in a person's mind is to be transmitted to the listener via speech. Convert the message into a sequence of words (mapping to words-phoneme sequences, prosodic pattern, intonation …) Neuromuscular commands perform articulatory mapping to control vocal cords, lips, jaw, tongue and velum. Production of the sound sequence as output.

Production of the sound sequence as output.
The Speech Chain (Human and Computer)

Message comprehension (not clear yet either)
The Speech Chain (Human and Computer)

Speech Generation
1) A semantic message in a person's mind is to be transmitted to the listener via speech
2) Convert the message into a sequence of words (mapping to words - phoneme sequences, prosodic pattern, intonation …)
3) Neuromuscular commands perform articulatory mapping to control vocal cords, lips, jaw, tongue and velum
4) Production of the sound sequence as output

Speech Understanding
5) Signal is passed to the cochlea in the inner ear which performs frequency analysis as a filter bank
6) A neural transduction process converts the spectral signal into activity signals in the auditory nerve
7) Neural signal is mapped into the language system (not clear)
8) Message comprehension (not clear)
The Structure of Spoken Language

- Speech Signals are composed of analog sound patterns
- The sound patterns serve as the basis for a symbolic, discrete representation of the spoken language - phonemes, syllables, words
- Production and representation of these sounds are governed by syntax and semantics
Overview

• **Spoken Language Structure**
  – The Speech Chain
  – The Structure of Spoken Language

  – **Sound**
    – What Is Sound?
    – Wavelength, Amplitude, Frequency of a Wave
    – Measuring the Intensity of Sound
    – Examples for Sound Levels in Decibel
    – Pick up the Sound Pressure Wave (Microphones)

• **Speech Perception**
  – Physiology of the Ear
  – Biology of Perception
  – Physical versus Perceptual Attributes
  – Non-uniform Equal Loudness
What Is Sound?

• Sound is a **pressure wave** which is created by a vibrating object.

• This vibrations set particles in the **surrounding medium** (typical air) in vibrational motion, thus **transporting energy** through the medium.

• Since the particles are moving in **parallel** direction to the wave movement, the sound wave is referred to as a **longitudinal wave**.

• The result of longitudinal waves is the creation of **compressions** and **rarefactions** within the air.

Source: media.allrefer.com

Longitudinal wave : Längswelle
Compression : Verdichtung
Rarefaction : Verdünnung

![Diagram of longitudinal wave and particle movements](media.allrefer.com)
What Is Sound?

The result of longitudinal waves is the creation of **compressions** and **rarefactions** within the air.

The particles do not move down the way with the wave but **oscillate back and forth** about their individual equilibrium position.

The alternating configuration of C and R of particles is described by the graph of a **sine wave** (C~crests, R~troughs)

The speed of a sound pressure wave in air is \(331.5 + 0.6T_c\) \(\text{m/s}\)  
\((T_c\) temperature in Celsius\)

For comparison: The speed of electromagnetic waves (from radio waves to X-rays) is \(~300.000\ \text{km/s}!\)
Wavelength, Amplitude, Frequency of a Wave

The amount of work done to generate the energy that sets the particles in motion is reflected in the degree of displacement which is measured as the amplitude of a sound.

The frequency $f$ of a wave is measured as the number of complete back-and-forth vibrations of a particle of the medium per unit of time.

1 Hertz = 1 vibration/second

$f = 1/\text{Time}$

Depending on the medium, sound travels at some speed $c$ which defines the wavelength $l$: $l = c/f$
Measuring the Intensity of Sound

• The softest audible sound modulates the air pressure by \( \sim 10^{-5} \) Pascal \( (1 \text{Pascal} = 1 \text{N/m}^2=1 \text{ kg/ms}^2) \)

• The loudest (pain inflict) audible sound does it by \( 100 = 10^2 \) Pa

• Because of this wide range it is convenient to measure sound amplitude on a logarithmic scale in **Decibel [dB]** *(named in the honor of Alexander Graham Bell – 1 bel [B])*

• Bel expresses the *magnitude of a physical quantity relative to a specified reference level*, thus dB has no dimension / phys. unit

• \( L \) is the ratio of sounds \( P_0, P_1 \) then \( L_{\text{dB}} \) represents this in dB

\[
L_{\text{dB}} = 10 \log_{10} \left( \frac{P_1}{P_0} \right)
\]
Measuring the Intensity of Sound (2)

• In acoustics the reference level is typically set at the Threshold of Hearing (TOH) of an average human, i.e. $P_0 := 0\,\text{dB}$, $P_0 = 2 \times 10^{-5} \,\text{Pa}$

• Effective sound pressure is the root mean square of the deviation from the local ambient pressure caused by the sound wave

• Sound Pressure Level (SPL) or sound level $L_p$ is the logarithmic measure of the root mean square ($rms$) sound pressure of a sound relative to $P_0 = P_{\text{ref}}$

$$L_p = 10 \log_{10} \left( \frac{p_{\text{rms}}^2}{p_{\text{ref}}^2} \right) = 20 \log_{10} \left( \frac{p_{\text{rms}}}{p_{\text{ref}}} \right) \,\text{dB}$$

• Face-to-face speech conversation (1 feet away) is ~ 60 dB SPL
• Close-talking microphone ~ 1 Pa = 94 dB
### Examples for Sound Levels in Decibel

<table>
<thead>
<tr>
<th>Description</th>
<th>Level</th>
<th>Description</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>threshold of hearing</td>
<td>0dB</td>
<td>softest audible 1000 Hz sound</td>
<td>6dB</td>
</tr>
<tr>
<td>quiet living room</td>
<td>20dB</td>
<td>soft whispering</td>
<td>25dB</td>
</tr>
<tr>
<td>refrigerator</td>
<td>40dB</td>
<td>soft talking</td>
<td>50dB</td>
</tr>
<tr>
<td>normal conversation</td>
<td>60dB</td>
<td>busy city street noise</td>
<td>70dB</td>
</tr>
<tr>
<td>passing motorcycle</td>
<td>90dB</td>
<td>somebody shouting</td>
<td>100dB</td>
</tr>
<tr>
<td>pneumatic drill</td>
<td>100dB</td>
<td>helicopter</td>
<td>110dB</td>
</tr>
<tr>
<td>loud rock concert</td>
<td>110dB</td>
<td>air raid siren</td>
<td>130dB</td>
</tr>
<tr>
<td>pain threshold</td>
<td>120dB</td>
<td>gunshot</td>
<td>140dB</td>
</tr>
<tr>
<td>rocket launch</td>
<td>180dB</td>
<td>instant perforation of eardrum</td>
<td>160dB</td>
</tr>
</tbody>
</table>

1) Threshold of hearing: One-billionth of a centimeter of molecular motion

2) The most intense sound (without physical damage) is one trillion times more intense.
Pick up the Sound Pressure Wave

**Pressure:**
- Air pressure is measured in Pascal (1Pa = 1N/m²=1kg/ms²)
- The standard air pressure on earth's surface is 1013 mbar (= 1013 hPA = 101300 Pa)
- The softest audible sound modulates the pressure by 0.000001 Pa ⇒ i.e. the standard pressure is 100 billion times higher than the softest audible sound

**Challenges:**
- Sound pressure decreases linearly with distance from the sound source
- Speed of excited air molecules decreases with the square of distance ⇒ Problems for recording speech using a microphone, far-field
Microphones

- Microphones convert the *acoustic energy (sound wave)* into *electrical energy (energy wave in voltage)*
- Microphone = Mike = Mic
- Mics measure the *molecular speed* or the *sound pressure*
- Lots of different Microphone types according to:
  - Principle of Operation:
    - Dynamic mic (electromagnetic generation),
    - Condensor mic (capacitance change)
    - Piezoelectricity mic (some materials produce voltage when subjected to pressure)
  - Polar-patterns
    - Omnidirectional
    - Unidirectional
  - Close-speaking vs. far-field microphones
- Usage:
  - Most typical one in ASR is the cardioid condensor microphone (range from throw-aways to high fidelity quality)

(http://en.wikipedia.org/wiki/Microphone)
Condenser Microphones

- Capacitor has two plates with a voltage between them (requires battery or external phantom power)
- Power is also needed to amplify the internal signal
- One plate is made of light material and acts as diaphragm, which …
  - vibrates when struck by sound waves,
  - changing the distance between the plates and thus
  - changing the capacitance.

http://www.mediacollege.com/audio/microphones/condenser.html
Challenges for Speech Recognition

Traditional Speech Recognition:
• Capture the acoustic sound wave by microphone
  Transform signal into electrical energy
• Requirements and Challenges:
  – **Audibility:**
    • Speech needs to be perceivable by mic
      (no low voice or whispering, no silent speech)
  \[\Rightarrow \textbf{Interference}:\] Speech disturbs others
    (no speaking in libraries, theaters, meetings)
  \[\Rightarrow \textbf{Privacy}:\] Speech signal can be captured by others
    (no confidential phone calls in public places)
  – **Robustness:**
    • Signal is corrupted by noisy environment
      (difficult to recognize in restaurants, bars, cars)
Implications for Speech Recognition

Recent Microphone developments with huge impact to ASR:
- Bone/skin transmitted vibration applying the principles of a medical stethoscope
  - Nakajima et al. (Nara Institute of Science and Technology): Stethoscopic microphone, Eurospeech 2003
  - Mayur Technologies: Magnetostrictive material coupled to the teeth/bone to convert the bone vibration into a proportional electrical signal
  - Microsoft, ASRU 2003
  - CMU, ISL (Stan Jou), ICSLP 2004

Applications in ASR:
- Noise Robustness (environment, cross-talk, ...)
- Privacy

<table>
<thead>
<tr>
<th></th>
<th>Normal Speech</th>
<th>Low Voice</th>
<th>Murmur</th>
<th>Whisper</th>
<th>NAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vibration of Vocal Chords</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Intention of Communication</td>
<td>YES</td>
<td>YES (Limited)</td>
<td>NO (Monologue)</td>
<td>YES (Limited)</td>
<td>NO (Monologue)</td>
</tr>
</tbody>
</table>
Bone-conduction

- When we speak normally our body is a resonance box
- Skin and bones vibrate when we speak (try this!)
- Capture this vibration by so-called bone-conducting or skin-conducting microphones

Zheng et al.

- Whispered speech is defined as:
  - the articulated production of respiratory sound
  - with few or no vibration of the vocal-folds
  - produced by the motion of the articulator apparatus
  - transmitted through the soft tissue or bones of the head

Nakajima

Jou et al. / Intecs
Whispered Speech

- Simultaneous recording throat and close-talk mic
- Speaking styles: normal vs. whispered
- Noise condition: quiet vs. party-babble

![Simultaneous recording](image)

**Close-talk**

<table>
<thead>
<tr>
<th>kHz</th>
<th>kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Throat mic**

<table>
<thead>
<tr>
<th>kHz</th>
<th>kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Normal** vs. **Whispered**
Whispered Speech – Performance

- Baseline System: Hub-4 Broadcast News: 10% WER recognition performance, normal speech, close talking, quiet (F0-condition)
- Experiments based on only 4 speakers, per spk:
  - Training: 38 phonetically balanced utterances
  - Test: 12 Broadcast News utterances
- Apply baseline system “as-is” to throat mic data:
  - 99% WER whisper – 77% WER normal
- After various adaptation steps incl. signal transformation:
  - 32.9% WER whisper – 21.7% WER normal

S-C. Jou, Schultz, Waibel: Whispy Speech Recognition using Adapted Articulatory Features, ICASSP-2005
Silent Speech Interfaces

• We just looked at whispered speech recognition.
• That does not, however, solve all the issues with ASR.

➤ Let's go one step further: Silent Speech Interfaces
   – Recognize Speech when no sound can be captured at all
   – May be used for voiceless persons (laryngectomy), private communication, loud environments

• At CSL, we investigate using electromyography for Silent Speech Recognition
   – Muscle activity is recorded with surface electrodes
   – Efficient and portable
   – Current performance: ~2000 words vocabulary, 45 minutes of training data, 11.1% WER

➤ acceptable for communication, but still some way to go
Overview

• Spoken Language Structure
  – The Speech Chain
  – The Structure of Spoken Language
  – Sound
    – What Is Sound?
    – Wavelength, Amplitude, Frequency of a Wave
    – Measuring the Intensity of Sound
    – Examples for Sound Levels in Decibel
    – Pick up the Sound Pressure Wave (Microphones)

• Speech Perception
  – Physiology of the Ear
  – Biology of Perception
  – Physical versus Perceptual Attributes
  – Non-uniform Equal Loudness
Physiology of the Ear

Three basic parts:

The **outer ear** serves to collect and channel sound to the middle ear.

The **middle ear** serves to transform the energy of a sound wave into the internal vibrations of the bone structure and transform these vibrations into a compressional wave in the inner ear.

The **inner ear** serves to transform the energy of a compressional wave within the inner ear fluid into nerve impulses which can be transmitted to the brain.

Anvil = Amboss
Stirrup = Steigbügel
Hammer = Hammer
Cochlea = Cochlea, Schnecke
Biology of Perception (1)

- **Sound waves** are guided from the *outer ear* to the *middle ear*, there it makes the *eardrum* move.

- A mechanical transducer (*hammer, anvil, stirrup*) adjacent to the eardrum’s opposite side converts the sound waves to *mechanical vibrations* on the *oval window*, the entrance to the cochlea.

- The *cochlea* is a spiral tube (3.5 cm long that coils about 2.6 times) filled with fluid in which standing waves are excited.
**Biology of Perception (2)**

- The **waves of the fluid** make the *cochlear filters* swing along.
- The *cochlear filters* are attached to the basilar membrane which responds to different **frequencies** at different locations.
- The **movement** of the *filters* is transferred to the brain along the *cochlear nerve*.

![Diagram of Cochlea](image)
Physical versus Perceptual Attributes

- Basic distinction between **perception of a sound** and its measurable **physical properties**:

  (Below listed items have a strong correlation but the connection is complex because other physical properties may affect the perception in complex ways.)

<table>
<thead>
<tr>
<th>Physical Quantity</th>
<th>Perceptual Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity</td>
<td>Loudness</td>
</tr>
<tr>
<td>Fundamental frequency</td>
<td>Pitch (Tonhöhe)</td>
</tr>
<tr>
<td>Spectral shape</td>
<td>Timbre (Klangfarbe)</td>
</tr>
<tr>
<td>Onset/offset time</td>
<td>Timing</td>
</tr>
<tr>
<td>Phase differences</td>
<td>Location</td>
</tr>
</tbody>
</table>
Non-uniform Equal Loudness

- One divergence between perceptual and physical quality is the non-uniform equal loudness: the sensitivity of the ear varies with frequency (Loudness ∝ Intensity)

- Ear sensitivity varies with frequency and quality of sound

- Insensitive at low frequency with low-moderate intensity, maximum at ~4kHz (= resonance frequency at outer ear), another peak at ~13kHz (= 2nd resonance frequency)

  \[\text{→ Human speech: 4kHz-13KHz}\]

- ISO Graph, Fig 2.11 [XH]
Overview

- **Speech Production**
  - Human Speech Production
  - Fundamental Frequency
  - Voiced vs. Unvoiced Sounds
  - Articulatory Properties
  - Speech Sound Names
  - Consonants versus Vowels
Speech is produced by air-pressure waves emanating from the mouth and the nostrils.

The production is a combination of:

1. Air-pressure waves effecting the vocal folds at the glottis (Stimmritze) (Stimmbänder)

2. Modulation of these waves by shaping the vocal tract

- Sounds can be partitioned into subgroups based on:
  - articulatory properties
  - presence/absence of constrictions or obstructions

Description of anatomical parts involved in speech production:

- Nasa/Nostril (Nasenlöcher)
- Gaumen (Gaumen / harter Gaumen)
- Gaumensegel (Gaumensegel / weicher Gaumen)
- Gaumenzapfchen (Gaumenzapfchen)
- Zahn (Zahn- damm)
- Zungen (Zunge)
- Rachen (Rachen)
- Kehle (Kehle)
- Kehlkopf (Kehlkopf)
- Stimmbänder (Stimmbänder)
- Stimmritze (Stimmritze)
- Stimmfalten (Stimmfalten)
- Lippen (Lippen)
- Speiseröhre (Speiseröhre)
- Luftröhre (Luftröhre)
- Nasenöhle (Nasenhöhle)
- Nasenlöcher (Nasenlöcher)
- Kehldeckel (Kehldeckel)
- Schildknorpel (Schildknorpel)
- Trachea (Trachea)

Artikulationsapparat
**Englisch - Deutsch**

**Larynx - Kehlkopf:**
Trachea - Luftröhre
Esophagus - Speiseröhre
Glottis – Stimmritze

(Ritze zwischen den Stimmbändern und den Stellknorpeln)

**Throat – Kehle:**
Epiglottis – Kehldeckel
Pharynx - Rachen

**Mouth/oral cavity - Mund(höhle):**
Palatum – Gaumen / harter Gaumen
Alveolus – Zahndamm
Tongue - Zunge
Lip – Lippen
Teeth – Zähne

**Nasal cavity - Nasenöhle**
Velum - Gaumensegel /weicher Gaumen
Uvula - Zäpfchen
Nostril - Nasenloch
The time for a single open-close cycle depends on the stiffness and size of the vocal folds and the amount of air pressure.

This can be controlled (in some range) by a speaker to raise and lower the *pitch* of a voiced sound.

**Vocal fold cycling:**
(a) Closed vocal folds build a barrier for the air stream from lungs
(b) Air pressure under the barrier overcomes the resistance of the vocal fold closure and blows them apart
(c) Elasticity in tissue and muscles make them fall back into place
Fundamental Frequency (Grundfrequenz)

- We can measure the number of such cycles per second. It is called the **fundamental frequency** $F_0$.
- It sets the periodic baseline for all higher-frequency harmonics by the resonance cavities above.
- $F_0$ varies from 60Hz (large men) to 300Hz (children).
Voiced vs. Unvoiced Sounds

- The most fundamental distinction between sound types in speech is the **voiced/voiceless** distinction
  - Voiced sounds have a roughly regular pattern in time and frequency structure
  - Voiceless sounds do NOT have this

- **Mechanism:**
  If the vocal cords vibrate during articulation the sound is voiced

- **Waveform of the word “sees”** consisting of an
  - unvoiced consonant /s/,
  - vowel /iy/
  - voiced consonant /z/
Articulatory Properties

- The sounds can be further subdivided into subgroups based on certain **articulatory properties**
- These properties are derived from the **anatomy** of important articulators and the **places** where they touch the boundaries of the human vocal tract
- The gross components of the human’s speech production apparatus are lungs, tracheae, larynx, pharyngal cavity (throat), oral and nasal cavity
- Pharyngal + oral cavity are referred to as vocal tract
- Nasal cavity as nasal tract
- The main apparatus consists of lungs, vocal cords (larynx), velum (soft palate), hard palate, tongue, teeth, lips
Speech Sound Names

Speech sounds get their names from the Latin name of the articulators and places:

- **Nasal** sounds: through nose (velum down) (e.g. /n/)
- **Oral** sounds: through mouth (velum up)
- **Stops**: full oral closure
- **Fricatives**: partial oral closure (friction) (Reibelaut)
- **Approximants**: narrowing (no friction) (Näherungslaut, Luft kommt seitlich durch)
- **Labial**: from labium, lip(s) active (die Lippen betreffend, z.B. /b/)
- **Dental**: from dentes, teeth active (die Zahn­lade betreffend)
- **Alveolar**: Alveoles, teeth ridge active
- **Palatal**: Palate, hard palate active (Gaumen)
- **Velar**: Velum, soft palate active (Gaumensegellaut)
- **Glottal**: Glottis, vocal cords active (Stimmbänder geschlossen, z.B. Be-amte)
Consonants versus Vowels

• In most of the world’s languages, the inventory of sounds can be divided into two basic classes:

• **Consonants**: articulated in the presence of constrictions in the throat or obstructions in the mouth as we speak

• **Vowels**: articulated without major constrictions and obstructions

(i.d.Regelohne Behinderung des Luftstroms im Vokaltrakt)

• **Facts** about Vowels and Consonants important to Speech Recognition:
  – The average duration of a vowel is much longer than that of a consonant
  – Vowels carry most of the energy of a signal, Consonants are weak
  – Thus vowels have a big impact on the recognition performance but consonants are difficult to recognize (they often look like silence)
  – For (English) text, the rule is just the opposite:
Thanks for your interest!